UNIVERSITY OF GONDAR
COLLEGE OF MEDICINE AND HEALTH SCIENCE
INSTITUTE OF PUBLIC HEALTH

PREDICTION OF MALARIA SPECIES MORBIDITY USING DATA MINING TECHNIQUE: THE CASE OF CHEWAKA HEALTH CENTER ILU ABA BORA ZONE, OROMIA NATIONAL REGIONAL STATE, SOUTH WEST ETHIOPIA, 2012.

By :- Dereje Oljira

Name of advisor(s):-

Mr. Kasahun Alemu(BSc, MPH)

Mr. Atinikut Alamirrew(BSc, MPH/HI)

A THESIS SUBMITTED TO THE INSTITUTE OF PUBLIC HEALTH, COLLEGE OF MEDICINE AND HEATLH SCIENCES, UNIVERSITY OF GONDAR IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PUBLIC HEALTH IN HEALTH INFORMATICS.

June, 2012

Gondar, Ethiopia

UNIVERSITY OF GONDAR

COLLEGE OF MEDICINE AND HEALTH SCIENCE

INSTITUTE OF PUBLIC HEALTH

PREDICTION OF MALARIA SPECIES MORBIDITY USING DATA MINING TECHNIQUE: THE CASE OF CHEWAKA HEALTH CENTER ILU ABA BORA, SOUTH WEST ETHIOPIA.

By: Dereje Oljira

Address:

Tell: +251-917-818-339

P. O .Box: 14 Bedele Town Health Office

Email: dodbau3687@gmail.com

Approved by the examining board

_____ _____

Head, Institute of public health

Advisors: Kasahun Alemu (BSc, MPH,) _____

Atinikut Alamirrew (BSc, MPH/HI) _____

_____ _____

Examiner

June, 2012

Gondar, Ethiopia

Dedication

I would like to dedicate this paper to my father and mother, Ato Oljira Donacho Tokon and W/ro Bajige Bachara Lamu, who have always been there and supported me,

Acknowledgement

I would like to forward my deepest gratitude to my advisors Mr. Kasahun Alemu and Mr. Atinikut Alamirrew for their constructive advice and valuable comments throughout the whole processes of the development of this thesis.

Very special thank goes to W/ro Bashatu Abebe ,Ato Desaleng Kebede, Ato Adamu Biru, Mr. Mulatu Ayana , Mr. Ashenafi Gadisa,my Mother w/ro Bejige Abachara,Ato Abera Shibiru,Ato Amanuel Namomsa, Ato Dabala Abdisa and Mr. Solomon Meseret for their support and also Ilu Aba Bora Zonal health department and Chewaka Woreda health office.

Acronyms

CART=Classification and Regression Tree

CRISP-DM= Cross-Industry Standard Process for Data Mining

CSA= Central Statistics Agency

NMA=National Meteorology Agency

WEKA=Waikato Environment for Knowledge Analysis

WHO-World Health Organization

Table of contents

Lists of Tables

List of Figures

Lists of Annexes

Abstract

Introduction: - Malaria is continues to be a leading cause of morbidity and mortality worldwide. In Ethiopia, transmission is unstable and seasonal, with occasional devastating epidemics. In health facilities morbidity data can be used for prediction of occurrence of disease and can help in decision making using data mining techniques.

Objective: - The main objective of this study was to predict malaria species morbidity from malaria data by using data mining technique in Chewaka Health Center, South-West Ethiopia, 2012.

Methods: Institution based retrospective record review study was conducted. All malaria positive data in Chewaka health center from September 01, 2007 to December 30, 2011 was collected from manual records in to new format prepared for the study purpose and the data was integrated with meteorology data of nearby meteorology station. Data quality was assured by using cross-check up of data collected with manual available record. A total of 5077 records were used and data analysis was done by using WEKA classification decision tree J48 and neural network algorisms with two modes, 10 fold cross-validation and 90%-10% percentage split for training-testing modes.

Results: Of 5077 records in dataset class attribute accounts *P. faliciparum* 2745(54.1%), *P. vivax* 2258(44.5%) and mixed 74 (1.5%). Prediction model developed by Decision tree J48 algorism with 90%-10% training-test mode was scored the highest accuracy and selected as best model. The model predicted correctly 86.22 % *P. faliciparum*, *P. vivax* 86.14 %and as mixed 99.4% species in their class, over all predictive accuracy was 90.5%. Mean monthly relative humidity, mean monthly maximum temperature, total monthly rainfall, Age and address were selected by the models as best predictive attributes for malaria species.

Conclusion and recommendation:-This study showed malaria morbidity to be predicted by mean monthly relative humidity, mean monthly maximum temperature, total monthly rainfall, Age and address. Using meteorology and morbidity data mining that can help for effective decision making on malaria prevention and early warning of epidemics is recommended.

1 INTRODUCTION

1.1 Statement of the problem

Malaria is one of leading cause of morbidity and mortality worldwide. According to the World Malaria Report 2011, malaria accounts for 98.5% of the deaths in Africa. This shows an estimated 216 million cases of acute malaria occur each year globally and 90 percent of the 2 million malaria deaths occur in Africa, mostly in young children(1,2).

In Ethiopia approximately 4-5 million cases of malaria are reported annually and the disease is prevalent in 75% of the country, putting over 50 million (68%) people at risk. It contributes up to 20% of under-five death, causes 70,000 deaths each year and accountant for 17% of outpatient visits to health institutions, 15% of admissions and 29% of inpatient deaths (3,4).

In Oromia three quarters of the region is considered malarias, accounting for over 17 million persons at risk of infection. There are an estimated 1.5 to 2 million clinical cases per year, accounting for 20-35% of outpatient consultations, 16% of hospital admissions, and 18-30% of hospital deaths in the region(4,5).

Malaria transmission unstable and largely unpredictable with occasional devastating epidemics during major harvesting season with serious consequences for the subsistence Economy of Ethiopia's countryside, and for the nation in general. The transmission patterns and intensity vary greatly due to the large diversity in altitude, rainfall, and population movement; areas below 2,000 meters are considered to be malarious (or potentially malarious)(3).

In Ethiopia there is massive data in health institution on malaria specially, in endemic areas such in chewaka district. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Now a days using morbidity data for prediction of future occurrences of particular problem is becoming introduced in modern technologies by using data mining techniques(6).

Data mining is the task of discovering interesting patterns from large amounts of data and process for different perspectives and summarizing it into useful information. It is possible to identify trends within data that go beyond simple analysis (7). In practice, data mining can accomplish about six common tasks namely; classification, estimation, prediction, association, clustering, and description(8).

Prediction is arguably the strongest goal of data mining. In predictive modeling one identifies patterns found in the data to predict future values. In health care system especially in disease such malaria in which transmission is unstable and unpredictable prediction model is important to generate hidden knowledge and in developing prediction of the future of the disease to support prevention and control activities(8).

This study aimed to predict malaria morbidity from malaria data in chewaka health center from the past three year's malaria data by using data mining technique. The mining task can help to find out prediction model that can show as the future occurrence of malaria based on the past morbidity data in the health center.

1.2 Literature Review

1.2.1 Data mining Techniques

Data mining is the task of discovering interesting patterns from large amounts of data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the ability to identify key attributes of business processes and target opportunities. In general, data mining tasks can be classified into two categories: descriptive and predictive(9, 10).

Descriptive mining tasks characterize the general properties of the data in the database whereas Predictive mining tasks perform inference on the current data inorder to make predictions. In practice, data mining can accomplish about six common tasks namely; classification, estimation, prediction, association, clustering, and description. A single data mining tool or technique is not equally applicable to all the above-mentioned tasks. Based on the nature of the problem under consideration and its proximity to the main divisions of data mining tasks; one need to choose the appropriate techniques among the numerous data mining technique(10).

Prediction is arguably the strongest goal of data mining and one can identify patterns found in the data to predict future values. Predictive modeling consists of several types of models: Classification models, regression models, and artificial intelligent models. Classification and regression are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends(11).

Classification methods create classes by examining already classified cases and inductively finding the pattern (or rule) typical to each class. Data mining uses machine learning methods using decision trees to classify objects based on a dependent variable. Regression models are the leading predictive models. The most common regression models are linear regression models, for modeling continuous response, and logistic regression models, for modeling discrete choice response (10).

The other types of predictive model are Artificial intelligence based models. The leading models in this category are neural networks models. It is biologically inspired model made up of a collection of processing units called neurons, connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. The most important feature of neural network model is it offers a means for large and complex problems with several independent variables(10).

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks(11).

Prediction encompasses the identification of distribution trends based on the available data. Classification and prediction may need to be preceded by relevancy analysis which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded(10).

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. Classification and prediction methods can be compared and evaluated according to their predictive accuracy, Speed, robustness, scalability and interpretability(10).

1.2.2 Malaria morbidity prediction

Finding from 2007 national malaria indicator survey in Ethiopia shows malaria is a seasonal disease and factors such as altitude, rainfall, and population movement are some of the factors for variability of the disease. It was also observed that areas below 2,000 meters above sea level are considered as non-malarias(12).

The role of malaria morbidity data is a great domain in analysis of change of trend and strategy of prevention and control of the disease. Results from India and Australia indicates morbidity data can be used for prediction of occurrence of vector born disease and can help in decision making in clinical medicine by using data mining techniques(13).

Study conducted in India revealed the application of data mining tools called CART (classification and Regression Tree) result in malaria prediction was meteorological variables such as maximum temperature, minimum temperature, rainfall, relative humidity, number of rainy days and month were ranked by CART. The result of this study were displayed in if rule model. The rule was “if rainfall=147.565, relative humidity<=89.305, minimum temperature<=3.5^{o}c and month=April, August, February, January, July, June March then malaria prevalence in the area will be 0.257817” and the other rule was “if rainfall>533.55Relative humidity>87.75 and maximum temperature >37 ^{o} c and month=June, July, August then malaria prevalence will be =24.9901(14).

In study conducted to predict malaria vectors by using data mining tools in Austral also indicated that rule-set prediction were developed from environmental variables and topographic parameters. Environmental attributes were associated with records of species were identified with the ranking procedures of decision tree software packages(15, 16).

The study conducted for Management of filariasis using prediction rules derived from data mining from India showed that Predictor variable (maximum temperature, minimum temperature, rain fall, relative humidity, wind speed, house type) are ranked by CART according to their influence on the target variable (month) for forecasting vector (Mosquito) densities in forthcoming seasons(17).

Including Ethiopia treatment records of patients in many countries of our world used for reporting and storage instead for source of knowledge discovery. But study done on prediction of heart attack using data mining technique indicated that treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health problem. Using medical profiles such as age, sex, blood pressure and sugar it can predict the likelihood of patients getting a heart disease(8). It is also possible to find out relationship of different disease with hidden cause, relationships and patterns(18).

Study conducted on epidemiology of malaria transmission with meteorological data, environmental remote sensing and neural network analysis in Thailand indicated that the average training accuracy of model is depend on data from malaria morbidity, environmental parameters and changes in meteorology records. The training and testing accuracy of urban and rural residences decrease when temperature parameter is removed the input (74±9% and 62±12%). When another hidden node of neural network is included more complex geometries can be introduced to assure better classification. The average testing accuracy of weighted malaria case was 53%(19).

Study conducted on model variation in predicting incidence of plasmodium malaria using morbidity and meteorology data from south Ethiopia indicated seasonal variation from different local at altitudes above 1742 meters, monthly rainfall, minimum and maximum temperature was able to predict incidence of the plasmodium malaria and relative humidity was not able to predict plasmodium malaria incidence(20). Study conducted in china indicated that the minimum threshold temperature for parasite development of *plasmodium faliciparum* and *vivax* are 18 c^o and 15 c^o respectively. (Lower than 16c^o and higher than 30c^o has negative effect on the parasites)(21).

1.3 Justification of the Study

Chewaka Woreda is one of the malarious Woreda in Ilu Aba Bora Zone in Oromia, having 27 rural Kebele and one urban kebele in which all the settlements' are malarias.

Malaria always occurs in unpredictable situation in the Woreda. It is the leading cause of admission and Out Patient Department visit and more than 15,000 malaria records (from chewaka district health office five year malaria report) available in the health Center as well in Woreda health office that can be used for decision making and prediction of future occurrence of life-threatening disease of the area population (malaria).

This study was intended to predict malaria morbidity from malaria data by using data mining technique in chewaka Health center that can support decision making in services provision and prevention and control of malaria in the study area. It is useful that finding of this research will give a great input for the study area and in field of malaria prevention and control by prediction that help for early warning . It also used for researchers who will conduct further study in the study area on a related topic.As much as our literature review is concerned no such study was conducted on the topic in special situation of newly settlement area in Ethiopia.

2 OBJECTIVES

2.1 General Objective

The general objective of the study was to predict malaria species morbidity by using data mining techniques in chewaka health center Ilu Aba Bora Zone, Oromia Region, South west Ethiopia, 2012 G.C.

2.2 Specific Objectives

- To prepare quality dataset.
- To identify an appropriate predictive attributes from the data.
- To select sutable classification algorithm.
- To develop malaria morbidity prediction model.
- To evaluate the performance of the model.

3 METHODS AND MATERIALS

3.1 Study Design

Institution based retrospective record review study was used and Cross-Industry Standard Process for Data Mining (CRISP-DM) was followed to undertake the mining task.

3.2 Study area

The study was conducted in Chewaka Health Center, Chewaka district, Ilu Aba Bora zone, Oromia national regional state, south West Ethiopia. The district is found at 566 km from Addis Ababa in the South West direction. The district has a total of 64,564(63,254 rural and 1,310 urban) from all population 34,149 male and 30,115 female) populations residing in 28 kebeles(22).The climatic condition of the district is 'Kola'(low land). The district accounts 54,220 hectares of land area which is all the land masses are suitable for settlement.

The annual temperature and rainfall is 26-37c^0 and 1000-1200mm respectively. Its altitude ranges from 900-1400m above sea level. The district is known the successful settlement in Oromia and in "salit" production. The main rivers found in the district are Didesa and Dabana which are tributary of Abay River.

The health center is located in main town of chewaka district (Ilu-Harar) and there are five health posts under the health center having 10 health Extension workers two in each health posts namely Terkanfata-Misoma, Shimal-Toke, Ilu-Harar, Demeksa and Duki. Malaria is the top cause of morbidity in the health center.

3.3 Source population

In this study all malaria positive cases in chewaka health center from September 07, 2007 to December 31, 2011. were used as source population.

3.3.1 Inclusion criteria

All malaria positive case those registered and confirmed by laboratory test were included.

3.3.2 Exclusion criteria

None readable, record containing incomplete in special attributes of this study was excluded (i.e species).

3.4 Sampling

The project was concentrated to records from September 01, 2007 to December 30, 2011.G.C. This records were selected based on intend to time proximity and easy for access to the documents.

3.5 Variables of the study

Dependent variable (class attribute):- malaria morbidity (malaria species)
Independent variables:

- ***Socio-demographic variables***:- Age , Sex, address (Kebele), Residence (urban, rural)
- ***Meteorology variables***: - Mean minimum monthly temperature, Mean maximum monthly temperature, Mean monthly rain fall, mean monthly humidity.
- ***Others variables***: - month, season of the year.

3.6 Operational Definition

❖ **Decision tree** is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions(9).

❖ **Neural network** is a collection of linear threshold units that can be trained to distinguish objects of different classes (11).

❖ **Classification** is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown(11).

❖ **Predictive accuracy** refers to the ability of the model to correctly predict the class label of new or previously unseen data(10).

3.7 Data collection Tools and procedures

Data was collected manually from the paper based records in the organization (chewaka Health Center) to new computerized format by data collection team consisting two diploma level in Information technology professional, two Nurses and one senior supervisor (health Officer) and data from meteorology station with respects to the study area with morbidity was integrated during data manipulation manually referring the month of the case with meteorology data of that month.

3.8 Data Quality Assurance

In all steps of the procedures of data collection and data cleaning records were cross-checked with manual documents in hard copy and soft copy through supervision. Two days training was given for data collectors and supervisors.

3.9 Data Processing and analysis

The morbidity data was collected by using format prepared for this study porpuse in a way that to track errors in Microsoft access 2007 and stored using Microsoft excel 2007.The total monthly rain fall, mean monthly temperature and mean monthly humidity was integrated with the records from morbidity.

Data processing and cleansing was imposed in order to make the data more suitable for the particular data mining software used in the study. This comprised accounting for missing data fields and the processed data was organized in a file format acceptable to the data analysis software. WEKA 3.7.4 version soft ware was used for data analysis.

WEKA was selected due that it is open source and has the facility to extract a random sample and then test the accuracy of the classifier on disjoint collection of cases under study. There are three options for partition the dataset in to training and test dataset: preparing distinct files for training dataset and testing data set, cross-validation with possibility of setting variety number of folds and percentage split.

In this study Prediction model was developed by using WEKA classification decision tree J48 and neural network algorism with two modes, 10 fold cross-validation and 90%-10% percentage split for training-testing modes. Based on selected model prediction rules were derived.

In the data analysis Cross-Industry standard Process for Data mining (CRISP-DM) was considered to care out the mining tasks. (figure1)

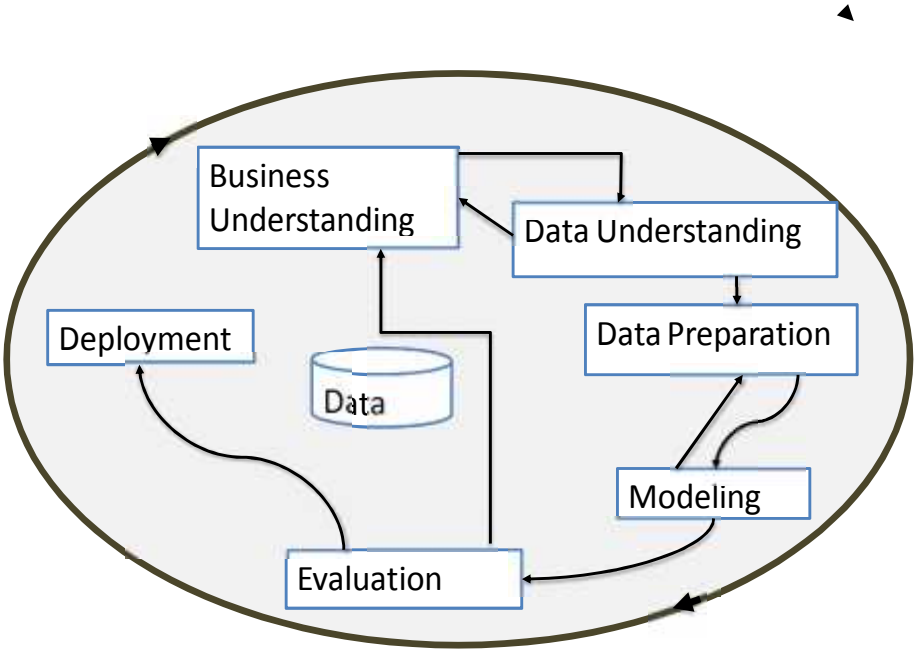

Figure 1: The Cross-Industry standard process for data Mining (CRISP-DM).

(Source: Fayyad, Piatetsky-shapiro and Smyth; 1996)(23).

The figure above indicates the knowledge discovery from data in CRISP-DM method that flows throughout six steps: Business understanding, data understanding, data preparation, modeling, evaluation and deployment. In this method moving back and forth in stage is possible(23).

3.10 Business Understanding

Understanding the business perspectives of a project and formulating data mining problem with clear objectives based on the knowledge on business problem is the first phase of data mining task(23). In our case, malaria is among vector born disease that transmitted by bite female anopheles mosquito. The transmission depends on different factors, such as rain fall, temperature, humidity, altitude and others(21).

Among the prevention and control activities; early treatment of cases, environmental management (such as removing of stagnant water), indoor insect side spray, using insect treated nets, and chemoprophylaxis(1, 2).The two species in the study area were *Plasmodium faliciparum* and *Plasmodium vivax*. Knowing the species can help in effective treatment and planning logistics for control.

In health care provision in Chewaka health center, patient registration is the first task that done at patient registration department. Then they come to physical examination in outpatient department, to confirm the problem lab investigation can be requested. For malaria similarly diagnosis done based on the physical diagnosis. In outpatient department there is registration for the diagnosis, again there is malaria diagnosis registration book in the laboratory department separately.

3.11 Data Understanding

The data used in this research project was taken from two government organization; malaria morbidity data from chewaka Health Center and the study area meteorology data were from Dedesa Diga meteorology station.

Malaria morbidity data: In this study two registration books were used: OPD patient registration that of before new HMIS registration started(having the patient information with records like medical record number, age, sex, address, date of diagnosis, months, diagnosis and treatments) (annex 2) and malaria diagnosis laboratory registration book (having medical record number, age, sex, address, date of diagnosis, months, test result and species)(annex 3) The data was collected from the two registrations based on inclusion and exclusion criteria, accordingly 5077 records was used.

Meteorology data: In Meteorology Agency there is stations database. For this study purpose the respected station Didesa Diga station was selected and data on monthly rain fall, humidity, and temperature was used (annex 4).

3.12 Data preparation and organization for analysis

As discussed in data preprocessing section, raw data were stored in excel format. But WEKA (Waikato Environment for Knowledge Analysis) accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation-file-Format; i.e file name.arff). Since the raw data we have is in excel format the step was simply save as comma delimited (CSV) format. Then open the file with the word pad program and organized in attribute relation file format (annex 5).At the end the data was saved with unique file name and extension (i. e File name.arff).

3.11 Defining the target attribute

From our objective we concerned to predict malaria morbidity with the class attribute type or species that were classified as *P. vivax*, *P. faliciparum*, and *mixed*.

4 ETHICAL CONSIDERATION

Ethical clearance was obtained from the Ethical review board of University of Gondar and supporting letter was obtained from University of Gondar, from Ilu Aba Bora Zonal Health Department, Chewaka Woreda Health Office and national meteorology agency to the concerned organizations. The purpose and importance of the study was explained to the Health managers of the Health Center, permission to proceed was assured and then the data was collected and confidentiality of the information was secured.

5 DISSEMINATION AND DISTRIBUTION OF THE RESULTS

The results of the study will be presented to the institute of public health and it will be presented to those who are in need of these result and accordingly it will be advocated for those who can implement it, example to the Ilu Aba Bora zonal health department, Chewaka district health office and chewaka health center administration and for those partners who are in field of malaria prevention and control activities in order to help them in management and epidemic prevention and control.

6 RESULTS

6.1 Characteristics of the selected records

Total of 5077 records were used from 6000 available records in the selected sample period and the rest 923 records were excluded based on the inclusion and exclusion criteria. Of this 5077 records in the class attribute accounts, *P. faliciparum* 2745(54.1 %), *P. vivax* 2258(44.5%) and mixed 74 (1.5%), mean age was16.25 Year (SD±13.20), totally 56% were male the rest 44%female. (Table 1)

Table 1 Socio-demographic distribution of selected malaria positive records in chewaka health center from S eptember 7, 2007 to December 30, 2011.

Attributes		Number	percent
Sex			
	Male	2843	56.0
	Female	2234	44.0
Address			
	Gudure	2411	47.5
	Demeksa	530	10.4
	Duki	180	3.5
	Shimaltoke	832	16.4
	Tokuma harar	104	2.0
	Haro chewaka	303	6.0
	Tarkanfata	336	6.6
	Waltasis	104	2.0
	Dursitu	277	5.5
Residence			
	Urban	2411	47.5
	Rural	2666	52.5

As showed in table 2 below the mean of mean monthly minimum temperature 15.307 **c^o (**SD ± 1.58), mean of mean monthly maximum temperature 26.6 **c^o**(SD ± 2.331), mean of mean monthly humidity 72.469% (SD ± 9.455) and the mean of total monthly rain fall was 152.798mml(SD ± 137.992). Averagely the maximum temperature ranges from 26.6c^o to 37 c^o, minimum temperature ranges from 11.6 c^o to 19.6 c^o.

Table 2 Description of selected meteorology variables record of Chewaka health center catchment area from September, 7 2007 to December 30, 2011.

Attributes	Description			
	Maximum	Minimum	Mean	Standard division
Total monthly rain fall(mm)	514.9	0	152.798	137.992
Mean monthly minimum temperature(c^o)	19.6	11.6	15.307	1.587
Mean monthly maximum temperature(c^o)	37	26.6	31.485	2.331
Mean monthly Humidity (%)	83.3	49.6	72.469	9.455

More than half of cases occurred in autumn and summer (71.7%), 41.7%of *p. faliciparum* in autumn and 30.1% in summer, 32.8% of *p. vivax* in autumn and 38.4% in summer. From the total cases, 54.9% of *P. faliciparum* was from rural and 50.2% of *P. vivax* was from urban. In all species males (56%) and 0-4 age group (26%) were with high percentage (table 3).

Table 3 Malaria species distribution in ckewaka health center with other variables among the selected records from September 7, 2007 to December 30, 2011

Variables		**Malaria species**							
		P. faliciparum		*Mixed*		*P. vivax*		**Total**	
		Count	%	Count	%	Count	%	Count	%
Season of the year									
	Autumn	1144	41.7	52	70.3	744	32.9	1940	38.2
	Winter	395	14.4	7	9.5	2.86	12.7	688	13.6
	Spring	380	13.8	6	8.1	361	16	747	14.7
	Summer	826	30.1	9	12.2	867	38.4	1702	33.5
Residence									
	Rural	1507	54.9	34	45.9	1125	49.8	2666	52.5
	Urban	1238	45.1	40	54.1	1133	50.2	2411	47.5
Sex									
	Female	1201	43.8	29	39.2	1004	44.5	2234	44
	Male	1544	56.2	45	60.8	1254	55.5	2843	56
Age category									
	0-4	783	28.5	12	16.2	571	25.3	1366	26.9
	5-9	358	13	15	20.3.	268	11.9	641	12.6
	10-14	267	9.7	8	10.8	287	12.7	562	11.1
	15-19	196	7.1	6	8.1	158	7	360	7.1
	20-24	383	14	5	6.8	294	13	682	13.4
	25-29	300	10.9	6	8.1	272	12	578	11.4
	30-34	173	6.3	7	9.5	184	8.1	364	7.2
	35-39	103	3.8	2	2.7	64	2.8	169	3.3
	40-44	82	3	9	12.2	81	3.6	172	3.4
	45-49	39	1.4	2	2.7	33	1.5	74	1.5
	50-54	37	1.3	1	1.4	26	1.2	64	1.3
	55-59	4	0.1	0	0	6	0.3	10	0.2
	60-64	12	0.4	1	1.4	10	0.4	23	0.5
	64<	8	0.3	0	0	4	0.2	12	0.2
		2745	100	74	100	2258	100	5077	100

6.2 Model Building Experiments Using WEKA 3.7.4 SoftWare

Model building was done by using WEKA 3.6.4 soft ware decision tree J48 and neural network algorism. To select predictive model total of 16 experiments were conducted and the better accuracy model was selected.

Organization of the Experiments and results

In the first experiment we used 10 fold cross-validation and 90% for training and the rest 10% for testing model with decision tree J48 algorism. The tree result was tree with 518 size and 383 leaves. The confusion matrix was made from 5077 records implemented to the program 1761(64.15 %), 1411 (62.48 %) and 2 (2.70 %) were correctly classified as *P.falciparum, P. vivax* and mixed respectively. The rest 35.85 % *P.falciparum*, 37.5 %*P.vivax* and 97.3 % mixed was classified incorrectly (Table 4).

Table 4 Output result from the J48 decision tree that was done by the data as it is.

Actual	**Predicted**			**Total**	**Score**
	Faliciparum	Vivax	Mixed		**(accuracy rate)**
Faliciparum	1761	981	3	2745	64.15%
Vivax	846	1411	1	2258	62.48%
Mixed	46	26	2	74	2.70%
	2653	2418	6	5077	62.52%

As we have observed from the above table the accuracy of the model was very low. (62.52%) To get better accuracy model additional 16 experiments were done and the results were compared in their accuracy. (Table 5 and 6)

Four different record set were organized for the rest experiments:

a) Record set one (RS1):- using all the 13 attributes without balancing the class attribute. (Total= 5077, Falciparum 2745(54.1 %), vivax 2258(44.5%) and mixed 74 (1.5%))

b) Record set two (RS2):-using all the 13 attributes records after eleven levels balance the class attributes using synthetic minority over sampling technique (SMOTE) algorism in WEKA filter package based on our computer memory capacity and the balance among the classes of class attribute .(Total=29,484 ,Falciparum 10980 (37.2%), vivax 9032 (30.7%)and mixed 9472 (32.1%))

c) Record set four (RS3):- using all the 13 attributes without balancing the class attribute and applying the re-sampling technique in WEKA filter package. (Total= 5077, Falciparum 2725 (53.7%), vivax 2284(45%) and mixed 68(1.3%)).

d) Record set three(RS4):-using all the 13 attributes after balanced (in record set two above) and using the WEKA re-sampling techniques in weka filter package. (Total =29484, Falciparum10958 (37.2%), vivax 9044(30.7%) and mixed 9482(32.1%)).

For each record set two test modes were used:-

1. 10-fold cross-validation mode (mode 1 for our case) and
2. 90% percent split of each record was used for model building and the rest 10% for testing of performance (mode 2 for our case). Then the results of the experiments were summarized in table 5 and table 6 below.

Table 5: Input parameters and J48 decision tree output parameters resulted from the 8 different experiments with their ranks of accuracy.

Ext. no	Record set	Attributes		Test mode	Tree size	Number of leaves	Time taken	Accuracy	Rank
		Input	Output						
1	RS1	13	13	Mode1	518	383	0.52	62.52%	7
2	RS2	13	13	Mode 1	2608	1826	6.05	88%	4
3	RS3	13	13	Mode 1	1163	887	0.91	70.3%	5
4	RS4	13	13	Mode 1	2826	1979	5.94	90.37%	2
5	RS1	13	13	Mode 2	518	383	0.54	59.6%	8
6	RS2	13	13	Mode 2	2608	1826	5.15	88.74%	3
7	RS3	13	13	Mode 2	1163	887	0.6	67.9%	6
8	RS4	13	13	Mode 2	2826	1979	6.11	90.50%	1

The table above indicted that the highiest accuracy was achieved on experiment 8(90.5%). In both modes results done by balanced and resampled record set has the better accuracy.

Table 6 Input parameters and the resulting outputs from neural network of 8 experiments with their ranks of accuracy.

Expt. no	Record set	Number attributes	Test mode	Test nodes	Time taken(sec)	Accuracy	Rank
9	RS1	13	Mode 1	27	381.52	57.71%	8
10	RS2	13	Mode 1	27	2046.78	81.48%	4
11	RS3	13	Mode 1	27	368.78	67.74%	5
12	RS4	13	Mode 1	27	1759.96	82.43%	3
13	RS1	13	Mode 2	27	379.3	56.89%	7
14	RS2	13	Mode 2	27	2043.94	**83.31%**	**1**
15	RS3	13	Mode 2	27	393.99	63.39%	6
16	RS4	13	Mode 2	27	914.9	83.2%	2

The highest accuracy level was 83.31% that was done built by experiment 14.

When we compare the result from J48 with Neural network, J48 was better accuracy in both test modes (Experiment 8). Experiment 8 was selected and used for farther analysis. The model in the selected experiment (Experiment 8) was resulted with confusion matrix as presented in table below. (Table 7)

Table 7 Confusion Matrix for model built on experiment 8 using J48.

Actual	**Predicted**			**Total**	**Score**
	Falciparum	Vivax	Mixed		**(accuracy rate)**
Falciparum	932	147	2	1081	86.22%
Vivax	119	783	7	909	86.14%
Mixed	3	2	953	958	99.48%
Total	1054	932	962	2948	90.502%

The confusion Matrix was made from 2948 test records, the model predicted *P. faliciparum* 932(86.22%), *P. vivax* 783 (86.14%) and mixed 962(99.48%) species correctly. The rest 13.78% of *P. faliciparum*, 13.86 % *P. vivax* and 0.52% of mixed species predicted incorrectly. The overall accuracy of the model was 90.502%.The decision tree of the model was with 2826 tree size and 1979 leaves.

To select best predictive attributes from all attributes weka.attribute.Selection. Best First -D 1 -N 5 algorism were used for all the experiment record set (RS1, RS2, RS3 and RS4) with the two test modes in our experiment were done. The result of the ranked attributes based on the information gain indicates that mean monthly relative humidity, mean monthly maximum temperature, total monthly rainfall, mean monthly minimum temperature, Age and address were selected by different testes as their frequency listed as best predictive variables for malaria species. This result was similar with the variables floated to the top of selected tree Experiment (Experiment 8) for analysis.

Using 90%-10% training-test instances split, removing other attributes (other than the selected above), experiment 8(the selected experiment) repeated by modifying the some parameters. The result showed tree size of 1873 and 979 of leave. The confusion matrix presented in table below (Table 8).

Table 8 Confusion Matrix for experiment 8 that built after some parameters modified.

Actual	**Predicted**			**Total**	**Score**
	Falciparum	Vivax	Mixed		**(accuracy rate)**
Falciparum	861	207	13	1081	79.64
Vivax	165	731	13	909	80.4
Mixed	5	4	949	958	99.1
Total	1031	942	975	2948	86.2

The confusion matrix indicated low accuracy when compared with the experiment done without modification. (86.2 %< 90.5%) From the all comparison above, experiment 8 that was done without modification achieved better accuracy (90.5%) and it was selected for rule generation. Due to large size of the tree, some part of the tree displayed in figure below (figure 3) the full tree annexed for reference. (Annex 1)

```
J48 pruned tree
------------------
Relative humudity <= 81.993593
|   Maxmumtemprature <= 33.285215
|   |   Adress = gudure
|   |   |   Maxmumtemprature <= 28.343425
|   |   |   |   Agecat = 20-24
|   |   |   |   |   Age <= 21.909829
|   |   |   |   |   |   Age <= 20.957225
|   |   |   |   |   |   |   Age <= 19.975773: faliciparum (7.0)
|   |   |   |   |   |   |   Rainfall > 1.351259
|   |   |   |   |   |   |   |   Minmum temprature <= 13.749684: faliciparum (233.0/3.0)
|   |   |   |   |   |   |   |   Minmum temprature > 13.749684
|   |   |   |   |   |   |   |   |   Agecat = 20-24: vivax (5.0)
|   |   |   |   |   |   |   |   |   Agecat = 45-49: faliciparum (11.0)
|   |   |   |   |   |   |   Maxmumtemprature <= 33.046341
|   |   |   |   |   |   |   |   Recidence = Urban
|   |   |   |   |   |   |   |   |   Rainfall <= 276.852629
|   |   |   |   |   |   |   |   |   |   Maxmumtemprature <= 31.895373
|   |   |   |   |   |   |   |   |   |   |   Sex = female
|   |   |   |   |   |   |   |   |   |   |   |   Rainfall <= 52.733103: faliciparum (84.0/8.0)
|   |   |   |   |   |   |   |   |   |   |   |   Rainfall > 52.733103
|   |   |   |   |   |   |   |   |   |   |   |   |   Maxmumtemprature <= 31.247545
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Age <= 1.036243
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Maxmumtemprature <= 30.14933: vivax (14.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Maxmumtemprature > 30.14933: faliciparum (65.0/28.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Rainfall > 23.877247: faliciparum (36.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Minmum temprature > 15.549383: vivax (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Seasonoftheyear = Winter
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Age <= 19.504904: faliciparum (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Age > 19.504904: vivax (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Seasonoftheyear = Spring: mixed (211.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Seasonoftheyear = Summer
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Maxmumtemprature <= 30.844012: vivax (6.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Maxmumtemprature > 30.844012: faliciparum (8.0)
Relative humudity > 81.993593
|   Seasonoftheyear = Autumn: vivax (2.0)
|   Seasonoftheyear = Spring: mixed (117.0)
|   Seasonoftheyear = Summer
|   |   Relative humudity <= 82.3
|   |   |   Agecat = 20-24
|   |   |   |   Month = Augest
|   |   |   |   |   Adress = gudure
|   |   |   |   |   |   Sex = female
|   |   |   |   |   |   |   Rainfall <= 262.306284
|   |   |   |   |   |   |   |   Age <= 19.642176: faliciparum (3.0)
```

Figure 2 Partial view of selected model tree that built by decision tree J48 algorism (experiment 8).

The selected model was tested with 2948 instances (10% of the total instances) and the confusion matrix presented in table 8 above. The measures of predictive performance resulted in weighted average precision of 0.905(0.884 for *P. faliciparum*, 0.84 for *P. vivax*, 0.991 for *mixed*) and recall of 0.905(0.862 for *P. faliciparum*, 0.861 for *P. vivax* and 0.995 for *mixed.)* (Table 9)

Table 9: Precision and recall accuracy of measures for model selected (Experiment 8)

Classes	Precision	Recall
Falciparum	0.884	0.862
Vivax	0.84	0.861
Mixed	0.991	0.995
Weighted Average	0.905	0.905

6.3 Some of the rules generated from the selected Decision tree

From decision tree it is possible to generate rules by simply combining the roots to the leaves nodes through the path in the tree(10). Accordingly the followings were some of the rules those can be generated from our model.

1. *‘‘If mean monthly relative humidity <=81.99% **and** mean monthly maximum temperature <=33.28C^{o} **and** total monthly Rain fall >1.35mm and mean monthly minimum temperature <=13.749 C^{o}**then P.faliciparum’’***

2. *‘‘If mean monthly relative humidity <=81.99% **and** mean monthly maximum temperature <=28.3425 C^{o}**and** Adress =Gudere **and** Age category 20-24 and age <=19.97 then **P.faliciparum** .’’*

3. *‘‘If mean monthly relative humidity >81.99% **and** mean monthly maximum temperature >28.2C^{o} **and** total monthly Rain fall> 259.1mm and season of the year Autumn then **P.vivax,** if season of the year **spring mixed** .’’*

4. *‘‘If mean monthly relative humidity <=59.3% **and** total monthly rain fall <0.04mm **and** mean monthly maximum temperature >33.29C^{o} **and** residence =Rural **and** Relative humidity >54.78 t**hen P.falciparum.’’***

5. *‘‘If mean monthly relative humidity <=51% and total monthly rain fall >0 and mean monthly maximum temperature <=35C^{o} then **P.vivax.’’***

7 DISCUSSION

This study was intended to predict malaria morbidity and two data mining algorisms (J48 and neural network) with two test modes and four different record sets were used. In both test nodes the four record sets (data without balancing class attribute records, data balancing class attribute records, data without balancing class attribute records but re-sampling and data balancing class attribute then re-sampling) were applied to compare predictive accuracy.

Model built by balanced record set was with the better accuracy when compared by unbalanced. Model built by balanced re-sampled record set was again better in accuracy than the other. The model with the test mode with 90%-10% training-testing achieved better accuracy in J48 algorism but low accuracy in neural network algorism. In all cases model built by J48 algorism was better in accuracy.

From different experiments Decision tree model developed by J48 algorism with 90%-10% training-test mode scored the highest accuracy of all and selected for our model. This model predicted correctly 86.22 % *P. faliciparum, P. vivax* 86.14 %and as mixed 99.4% species in their class and indicated 90.5% over all predictive accuracy.This result has better accuracy when compared with related study conducted in Thailand to predict malaria cases that had 53% accuracy[19].

In the selection of predictive attributes, mean monthly relative humidity, mean monthly maximum temperature, total monthly rainfall, Age and address were selected by the models as best predictive attributes for malaria species. This finding is similar with the study conducted in India by using other data mining algorism[14].Other related study conducted in south Ethiopia indicated similar finding with rain fall predictive of malaria morbidity but not relative humidity[20]. This difference may be due to place variation and tools difference.

The rule generated from the decision tree model, relative humidity <=81.99% and maximum temperature <=33.28C^{o} and total monthly Rain fall >1.35mm and minimum temperature <=13.749 C^{o} predicted *P. faliciparum*, and relative humidity <=51% and rain fall >0 and maximum temperature <=35C^{o} predicted *P. vivax*. Related Study conducted in China indicated that the minimum threshold temperature for parasite development of *P. faliciparum* and *P. vivax* are 18 C^{o} and 15 C^{o} respectively. In our case we found prediction of *P. faliciparum* at <=13.74 C^{o}[21].This result may be due to resistance of plasimodium specias to cold temperature.

8 STRENGTHS AND LIMITATIONS OF THE STUDY

Strength

- Frequent supervision and on spot checking was made by supervisors and principal investigator during data collection.

Limitation

- Lack of similar study for comparison
- Since the data was collected from manual hard copy using large data set was not done that can result in better accuracy.
- Due to poor data management health institutions only limited variables were addressed.
- Meteorology variables like soil moisture ever flow and wind speed was not included.

9 CONCLUSION AND RECOMMENDATION

Conclusion

In conclusion the result showed that malaria morbidity and meteorology data can be used to predict malaria morbidity with reasonable accuracy. Decision tree built by J48 with 90%-10% training-testing mode was found the best predictive model for plasmodium malaria. Among all attributes used meteorology variables mean monthly relative humidity, mean monthly maximum temperature, total monthly rainfall and morbidity variables age and address were selected by the models as best predictive attributes for malaria species. The result found in this study rely it was hidden in manual stored records in health institution and meteorology station that the role of data mining can be appreciated.

Recommendation

For health institution

- Improving their patient record management system since it can use for knowledge discovery.
- Using meteorology information's by integrating their health data they can improve decision to prevent malaria morbidity in early warning.
- Malaria may occur in the out of the Known trained in the area due to climate change. So Health institutions should consider meteorology information to make effective decision on prevention of malaria.

For researchers

- Further investigation needed in the field of health care data mining specially to investigate associations among other infectious diseases and malaria.

10 REFERENCES

1. WHO. World Malaria Report. Geneva: World Health Organization; 2010.

2. WHO. World Malaria Report. Geneva: World Health Organization; 2011.

3. Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz J. Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. Malaria Journal. 2004 Nov 12;3(1):41-52.

4. Alemu A, Abebe G, Tsegaye W, Golassa L. Climatic variables and malaria transmission dynamics in Jimma town, South West Ethiopia. Malaria Journal. 2011;4(30):1-11.

5. Shargie EB, Gebre T, Ngondi J, Graves PM, Mosher AW, Emerson PM, et al. Malaria prevalence and mosquito net coverage in Oromia and SNNPR regions of Ethiopia. BMC Public Health. 2008;8(1):321-32.

6. Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. Malaria Journal. 2010;9(1):251-5.

7. Oluwagbemi O., Ofoezie O, Uzoamaka, Nwinyi, Obinna. A Knowledge-Based Data Mining System for Diagnosing Malaria Related Cases in Healthcare Management. Egyptian Computer Science Journal. 2010;34(4):195-202.

8. Srinivas K, Kavihta Rani B, Dr.Govrdhan A. Application of data mining Techniques in Healthcare and prediction of Heart Attacks. Journal on Computer Science and Engineering. 2010;02(02):250–5.

9. Krzysztof J, Witold P, Roman S. Data mining methods for knowledge discovery. New Jersey A John Wiley and Sons Inc; 1998.

10. Jiawei H, Micheline K. Data mining: concepts and techniques. Simon Fraser University Morgan Kaufmann Publishers;2006.

11. Daniel. T L. Data mining methods and models. New Jersey A John Wiley and Sons Inc; 2006.

12. MOH. Ethiopia Malaria Indicator Survey Report. Addis Ababa: Ethiopian Minister of Health; 2007.

13. Murty US. Application of data mining techniques for the control of vector borne diseases. India: Morehouse College Indian Institute of Chemical Technology; 2006.

14. Murty US. Application of Correlation & Regression Tree (CART) for management of Malaria in Arunachal Pradesh, India - ISPUB. The Internet Journal of Tropical Medicine. 2008;5(1).

15. Sweeney AW, Beebe NW, Cooper NB, Bauer JT, Peterson AT. Environmental factors associated with distribution and range limits of malaria vector Anopheles farauti in Australia. Journal of Medical Entomology. 2006;43(5):1068–75.

16. Sweeney A., Beebe N., Cooper R. Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. Ecological modelling. 2007 May 9;203(3-4):375–86.

17. Duvvuri VRS., Kumarawsamy S, Kadiri M., Murty U. Management of filariasis using prediction rules derived from data mining. Biomedical Informatics. 2005;1(1):8–11.

18. Zbynek B, Hagai G. Data mining of the transcriptome of Plasmodium falciparum: the pentose phosphate pathway and ancillary processes. Malaria Journal. 2005;4(1):17-25.

19. Kianq R, Adimi F, Soika V, Niqro J, Sinqhasivanon P, Sirichaisinthop J, et al. Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. Geospat Health. 2006;1(1):71–84.

20. Loha E, Lindtjorn B. Model variations in predicting incidence of Plasmodium falciparum malaria using 1998-2007 morbidity and meteorological data from south Ethiopia. Malaria Journal. 2010;(9):166-71.

21. Grace A. Influence of Climate on Malaria in China. Penn McNair Research Journal. 2011; 3(1):1-25

22. CSA. Population and Housing Census of Ethiopia. Addis Ababa: Centeral Stastics Agency; 2007.

23. Fayyad U, Jiawei H, Evangelos S. Knowledge Discovery and Data Mining Towards a Unifying Framework. 1996.

11 ANNEXES

Annex 1 Decision tree of selected model built by J48 algorism

J48 pruned tree

Relative humudity <= 81.993593

| Maxmumtemprature <= 33.285215

| | Adress = gudure

| | | Maxmumtemprature <= 28.343425: vivax (408.0/157.0)

| | | | | | | Minmum temprature <= 13.749684: faliciparum (255.0/11.0)

| | | | | | | Minmum temprature > 13.749684: vivax (16.0/0)

| | | Maxmumtemprature > 28.343425

| | | | Relative humudity <= 59.734832

| | | | | Agestatus = Abovefive

| | | | | | Rainfall <= 1.351259: vivax (16.0/10.0)

| | | | | | Rainfall > 1.351259: faliciparum (244.0/8.0)

| | | | | Agestatus = Underfive

| | | | | | Age <= 0.95857: vivax (28.0)

| | | | | | Age > 0.95857: faliciparum (98.0/92.0)

| | | | Relative humudity > 59.734832

| | | | | Age <= 2.002567

| | | | | | Maxmumtemprature <= 33.046341: faliciparum (550.0/225.0)

| | | | | | Maxmumtemprature > 33.046341: mixed (106.0/11.0)

| | | | | Age > 2.002567

| | | | | | Agecat = 20-24

| | | | | | | Age <= 19.95962: mixed (211.0/67.0)

| | | | | | | Age > 19.95962

| | | | | | | | Rainfall <= 336.404549:faliciparum (661.0/286.0)

| | | | | | | | Rainfall > 336.404549 : vivax (65.0/41.0)

| | | | | | Agecat = 45-49

| | | | | | | Minmum temprature <= 15.644898: faliciparum (102.0/39.0)

| | | | | | | Minmum temprature > 15.644898: mixed (37.0/4.0)

| | | | | | Agecat = 0-4 : mixed (2260.0/334.0)

| | | | | | Agecat = 25-29

| | | | | | | Rainfall <= 61.4: vivax (258.0/170.0)

| | | | | | | Rainfall > 61.4

| | | | | | | | Rainfall <= 212.133158 :mixed (1831.0/268.0)

| | | | | | | | Rainfall > 212.133158: vivax (133.0/130.0)

| | | | | | Agecat = 15-19

| | | | | | | Recidence = Urban : faliciparum (377.0/180.0)

| | | | | | | Recidence = Rural: mixed (529.0/44.0)

| | | | | | Agecat = 5-9

| | | | | | | Maxmumtemprature <= 32.050451

| | | | | | | | Agestatus = Abovefive: mixed (2472.0/317.0)

| | | | | | | | Agestatus = Underfive: faliciparum (90.0/60.0)

| | | | | | | Maxmumtemprature > 32.050451: vivax (166.0/80.0)

| | | | | | Agecat = 30-34

| | | | | | | Rainfall <= 23.877247: mixed (430.0/96.0)

| | | | | | | Rainfall > 23.877247

| | | | | | | | Rainfall <= 103.659367 faliciparum(66.0/15.0)

| | | | | | | | Rainfall > 103.659367

| | | | | | | | | Minmum temprature <= 15.7: vivax (139.0/89.0)

| | | | | | | | | Minmum temprature > 15.7 faliciparum (59.0/20.0)

| | | | | | Agecat = 35-39

| | | | | | | Rainfall <= 1.839359: vivax (27.0)

| | | | | | | Rainfall > 1.839359: faliciparum (160.0/67.0)

| | | | | | Agecat = 10-14

| | | | | | | Age <= 12.004334 : faliciparum (303.0/277.0)

| | | | | | | Age > 12.004334: mixed (637.0/153.0)

| | | | | | Agecat = >64: faliciparum (73.0/23.0.0)

| | | | | | Agecat = 40-44 :mixed(760.0/170.0)

| | | | | | Agecat = 60-64: vivax (16.0/6.0)

| | | | | | Agecat = 55-59: faliciparum (3.0)

| | Adress = demeksa

| | | Rainfall <= 299.7 : faliciparum (637.0/490.0)

| | | Rainfall > 299.7: vivax (212.0/18.0)

| | Adress = duki : vivax (139.0/66.0)

| | Adress = shimaltoke : faliciparum (996.0/697.0)

| | Adress = tokumaharar : faliciparum (69.0/33.0)

| | Adress = harochewaka : mixed(600.0/214.0)

| | Adress = tarkanfata : faliciparum (387.0/148.0)

| | Adress = waltasis :faliciparum (30.0/26.0)

| | Adress = dursitu :faliciparum(209.0/121.0)

| Maxmumtemprature > 33.285215

| | Month = Octomber: mixed (94.0)

| | Month = November:mixed(48.0/14.0)

| | Month = December: mixed(185.0/7.0)

| | Month = April : faliciparum (1010.0/283.0)

| | Month = February : vivax (450.0/237.0)

| | Month = January :vivax (532.0/308.0)

| | Month = March : vivax(391.0/200.0)

| | Month = May : faliciparum (33.0/3.0)

| | | | Minmum temprature <= 16.244597: faliciparum(145.0/110.0)

| | | | Minmum temprature > 16.244597: vivax (133.0/29.0)

Relative humudity > 81.993593

| Seasonoftheyear = Autumn: vivax (2.0)

| Seasonoftheyear = Spring: mixed (117.0)

| | Relative humudity <= 82.3 : vivax (2060.0/1254.0)

| | Relative humudity > 82.3 : faliciparum(1123.0/243.0)

Annex 2: Lists of attributes that found in Outpatient registration book used in Chewaka health center September 7, 2007-December 30, 2011.

Attributes	Data type	Description
MRN	Nominal	Unique Identification of patient number
Year	Numeric	Year of the diagnosis
Month	Nominal	Month of diagnosis
Age	Numeric	Patient age
Sex	Nominal	Gender of the patient
Address	Nominal	Address of the patient
Residence	Nominal	Residence of the patient
Sign and symptom	Nominal	Sign and symptom of the patient
Diagnosis	Nominal	Diagnosis of the patient
Treatment	Nominal	Medication given for the patient

Annex 3: Lists of attributes that found in malaria diagnosis registration book used in Chewaka health center September 7, 2007-December 30, 2011

Attributes	Data type	Description
MRN	Nominal	Unique Identification of patient number
Year	Numeric	Year of the diagnosis
Month	Nominal	Month of diagnosis
Age	Numeric	Patient age
Sex	Nominal	Gender of the patient
Address	Nominal	Address of the patient
Residence	Nominal	Residence of the patient
Test Result	Nominal	The result of patient diagnosis
Type (species)	Nominal	Malaria species identified by the diagnosis

Annex 4: Lists of attributes collected from Meteorology agency that of Chewaka Health Center Catchment area meteorology variables from September 7, 2007 to December 30, 2011.

Attributes	Data type	Description
Station code	Nominal	Unique Identification of meteorology station
Year	Numeric	Year at which record taken
Month	Nominal	Year at which record taken from the reading parameter.
Station name	Nominal	Station name
Station location	Nominal	Location of the station
Mean monthly minimum temperature	Numeric	Average monthly minimum temperature of the study area.
Mean monthly maximum temperature	Numeric	Average monthly maximum temperature of the study area.
Mean monthly Relative humidity	Numeric	Average relative humidity of the study area.
Total monthly rain fall	Numeric	Sum total of monthly rain fall of the study area

Annex 4: Lists of attributes selected and used for building classification model with their description and possible values to predict malaria morbidity by data mining in chewaka health center, 2012.

SN	Attributes	Data type	Description	Possible values
1	Age	Numeric	Patient age	Number
2	Sex	Nominal	Gender of the patient	Male or Female
3	Residence	Nominal	Residence of the patient	Urban or Rural
4	Month	Nominal	Month of the diagnosis	All month of the year
5	Age category	Nominal	CSA age category (standard age category)	0-4,5-9,10-14,15-19,20-24,25-29,30-34,35-39,40-44,45-49,50-54,55-59,60-64, >64
6	Address	Nominal	Address of the patient (Kebele of the patient)	Name of kebeles
7	Season	Nominal	Season of the year	Autumn, Spring, Summer, Winter
8	Age status	Nominal	Age categories of the patient's	<5 or =>5
9	Mean minimum temperature	Numeric	Average monthly minimum temperature of the study area.	Number
10	Mean monthly maximum temperature	Numeric	Average monthly maximum temperature of the study area.	Number
11	Mean monthly Relative humidity	Numeric	Average relative humidity of the study area.	Number
12	Total monthly rain fall	Numeric	Sum total of monthly rain fall of the study area	Number
13	Type(species)	Nominal	Species of malaria	Falciparum, vivax or mixed

Annex 5: Data preparation format with attributes label and sampled data for WEKA soft ware

```
@Relation Malaria_morbidity

@attribute Age {real}
@attribute Agestatus {=>5',<5'}
@attribute Agecategory{=>64',0 4',5 9',10 14',15 19',20 14',25 29',30 34',35 39',40
44',45-49',50-54',55-59',60-64',5-9'}
@attribute Sex {male,female}
@attribute Adress {shimaltoke, gudure, waltasis,dursitu,demeksa, tarkanfata, duki,
tokumaharar,harochewaka}
@attribute Residunce {urban,rural}
@attribute monthofdiagnosis {April, December, February,March, July,June,Octomber,
January, September, Augest, May, November}
@attribute Rainfall{real}
@attribute Maxtemprature{real}
@attribute Mintemprature{real}
@attribute RelativeHumudity{real}
@attribute seasonoftheyear {Autumn,Spring,Summer,Winter}
@attribute Type {mixed, vivax, faliciparum}

@data
20,=>5',20 24',female,gudure,Urban,Octombor,49.2,29.6,15.5,74.3,Autumn,faliciparum
```

Annex 6 Lists of Experiments

1. Experiment1 by using RS1 , J48 and mode 1
2. Experiment2 by using RS2, J48 and mode1.
3. Experiment 3 by using RS3, J48 and mode 1
4. Experiment 4 by using RS4, J48 and mode 1
5. Experiment 5 by using RS1, J48 and mode 2
6. Experiment 6 by using RS2, J48 and mode 2
7. Experiment 7 by using RS3, J48 and mode 2
8. Experiment 8 by using RS4, J48 and mode 2
9. Experiment 9 by using RS1, neural network and mode 1

10. Experiment 10 by using RS2, neural network and mode 1
11. Experiment 11 by using RS3, neural network and mode 1
12. Experiment 12 by using RS4, neural network and mode 1
13. Experiment 13 by using RS1, neural network and mode 2
14. Experiment 14 by using RS2, neural network and mode 2
15. Experiment 15 by using RS3, neural network and mode 2
16. Experiment 16 by using RS4, neural network and mode 2

Annex 7: - Format for data collection

Sr. No	**Field Name (Attribute)**	**Set values (Data Type)**	**Description**
1	Code	Auto Number	0001 to 6000
2	Age	Number	Real
3	Sex	Text	(male=m, female=f)
4	Address	Text	(name of kebeles)
5	Date of diagnosis	Text	(Ethiopian calendar, dd/mm/yyyy)
6	Month of diagnosis	Text	Name of months
7	Type	Text	(pf ,pv) the common two
8	Other diagnosis	Text	Other disease with the case
9	Residence	Text	Urban=u, Rural=r
10	Type of medication	Text	Name of drugs
12	Mean monthly minimum temperature	Number	Real

13	Mean monthly maximum temperature	Number	Real
14	Mean monthly humidity	Number	Real
15	Mean monthly rain fall	Number	Real

Annex-8: Information Sheet

Title of the Research project:

Prediction of malaria morbidity using data mining technique: The case of Chewaka Health Center Ilu Aba Bora, south west Ethiopia.

Name of Principal Investigator: Dereje Oljira Donacho

Name of Advisors: 1. Mr. Kasahun Alemu (BSc, MPH)

2. Mr. Atinikut Alamirrew (BSc MPH/HI)

Name of Organization: University of Gondar, College of Medicine and Health Sciences, Institute of Public Health**.**

Name of the Sponsor: University of Gondar

Introduction

This information sheet is prepared with aim of explaining the research project that you are asked to join with group of research investigators. The research group includes one main investigator, four trained data collectors, one supervisor and two advisors from University of Gondar.

Purpose: The purpose of this research study is to predict malaria morbidity by using data mining technique in chewaka health center, chewaka district, Ilu Aba Bora Zone, Oromia region south west Ethiopia. The result of this study will be used to make recommendation for those who are responsible for prevention and control of malaria and create appropriate management of malaria prevention and control activities in warning of epidemics.

Procedure: This study uses institution based retrospective record mining in health center and meteorological records of the study area that will be obtained from nearby meteorology station. Permission will be processed from the University of Gondar, National meteorology agency, Ilu Aba Bora zone health department, then from the chewaka district health office and chewaka health centers. Data will be collected from registration after permission obtained from concerned administrators.

Risk and/or Discomfort: There is no any risk or discomfort that will be faced by giving of the data except searching of documents from the store in dedication of time to do so. Every piece of information will be kept confidentially. There is no any risk from the research project.

Benefits: There will be benefit for all the community in endemic area of malaria in which this research result can predict future occurrence ahead the time of morbidity for prevention and control.

Incentives/Payment for Participating: There is no incentive or payment to be gained by taking part in this project.

Confidentiality: The information Collected from this research project will be kept confidential. Information will be accessed by the researcher and research assistant only.

Persons to contact: This research project will be reviewed and approved by the ethical committee of the University of Gondar. If you want to know more information you can contact the committee through the address below. If you have any question you may contact the following individuals.

Investigator: Dereje Oljira Donacho

Advisors:

1. Mr. Kasahun Alemu (BSc, MPH)
2. Mr. Atinikut Alamirrew (BSc, MPH/HI)

Declaration

I, the undersigned, senior MPH student declare that this thesis is my original work in partial fulfillment of the requirement for the degree of Master of Public Health.

Name _____

Signature_____

Place of submission: Institute of Public Health, Collage of Medicine and Health Sciences, University of Gondar.

Date of Submission: _____

This thesis work has been submitted for examination with my /our approval as university advisors:

Name	**signature**
1. _____	_____
2. _____	_____